JP34

化学グラフからのマイニングにおける線形フラグメント記述子の生成と選択

(関学大理工)
○岡田孝*、山川眞透

1. はじめに

著者らは数年来，ドーパミン受容体に活性を有する化合物群の特徴解析を例題として，カスケードモデルやそれに基づくデータスケープ探索の技法を開発してきた [1, 2]. この方法においては，与えられた構造式から多数の線形フラグメントを切り出し，各化合物をそれらフラグメントの有無で特徴づけている．これまでの解析では出力されるルールの解釈が必ずしも容易でないため，ルールを支持する構造式群の視察に頼って特徴を認識した．しかし，この作業はディスプレイ上でスクロールしながら行うため，非常な注意力が必要であった.

そこで，これまでの線形フラグメント表現を改良し，ルールの可読性を向上させることとした．また，得られたフラグメント群には相関の高いものが多数存在する．そこで，名義変数間の相関係数を定義し，それに基づく属性選択のプロセスを組み込むことにより，探索空間の縮小を試みた．これにより，支持度の低いフラグメントを含む多数の記述子群によるマイニングが可能になると考えられる．本稿ではこれらの改良の概要と結果に与えた影響について述べる.

2. 線形フラグメント記述子の改良

構造式からのフラグメント記述子生成法は，先の報告[3] を発展させたものであり，下記の方法によっている.

1. 指定した種類の元素および結合両端の原子を起点として，最短 path 長が max-length 以内のすべての原子との間で線形フラグメントを取り出す.
2. この線形フラグメントをその構成元素と結合の種類，各原子の配位数と付随水素原子の有無，および分岐枝上の最初の原子により特徴づける.
3. 利用者の指定した詳細度に従い，これらの線形フラグメントを記述する.

詳細な記述形式を採用すると，多種類の記述子が生成され，個々の記述子を支持する事例数が少なくなる．その結果，偏った分布を示す記述子が多くなり，活性の識別に有効なものが減少する．反面，簡単な記述形式では，マイニング結果の解釈が困難となる.

これまでの解析においては，いくつかの記述形式を検討した結果，線形フラグメントの両端から２個の原子のみに配位数と付随水素原子の有無を記載した形式を採用してきた．なお，線形フラグメントの長さは原子数が 10 以内に限っている．例えば，C3H:C3-C-C4H-N3 は 3 配位の CH と C の芳香環（:は aromatic bond）に，C-C-N がつながった構造を示す. 3 番目の C には配位数と水素の有無は記載されない. また, 3 配位の N に H が記されていないことから，この N は 3 級アミンであることが分かる.

今回以下の各節に示す機能を付与することにより，記述子の改良を試みた.

2.1 水素結合フラグメントの生成

これまでの解析で，構造式群の視察から分子内水素結合 XH..Y が重要な役割を果たしていると考えられる場合があった．ここで，X, Y は通常 O, N など電子的に陰性の元素である．しかし，上記のフラグメント生成アルゴリズムは構造式のトポロジカルな情報のみを利用しているため，水素結合を含むフラグメントは生成されない．他方，物理化学的性質の推定に用いた MM-AM1-geo 法は，半経験的分子軌道法計算により，不十分ながらも分子の３次元構造を与えており，これから分子内水素結合の存在を推定することができる.

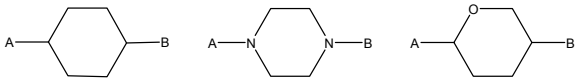
我々は，以下の条件を満たす場合に水素結合が存在すると見なし，それを表すフラグメント群 Xh.Y, V-Xh.Y, Xh.Y-W および V-Xh.Y-W を生成した．ここで，”h.”が水素結合を表し，V と W はそれぞれ X, Y に結合する原子を示す.

1. 原子 X が O, N, S または 4 配位の C で少なくとも１つの水素原子と結合していること.
2. 原子 Y が O, N, S, F, Cl, Br のいずれかであること.
3. XY 間距離が，Y が O, N, F のいずれかである時は 3.7 Å以下であり，それ以外では 4.2 Å以下であること.
4. 構造式上で X と Y の原子の間は，２原子以上で隔てられていること.

*okada-office@ksc.kwansei.ac.jp

2.2 フラグメントにおける中間表記の省略

分子の活性部分構造は往々にして，活性点に配位する活性部分とその間を結ぶリンカーで構成されている場合がある．リンカーは特定の立体位置に活性部分を配置することが役目であり，その原子種自体には意味の無い場合が多い．例えば，次の３種の構造で A と B を結ぶリンカーは，原子種は異なるがほぼ同一の位置に活性部分を配置することができる．



このような場合に，線形フラグメント中のすべての元素記号を記すと，３種の部分構造は互いに異なると認識され，支持事例が減少する．その結果どのフラグメントも採用されない場合がある．

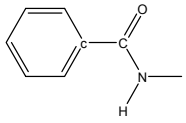
そこで，両端２原子以外については，その元素記号や結合記号を省略できるようにシステムを改造した．A, B が両端２原子を表すとすれば，上記の３構造はすべて下記のフラグメントで表される．

結合記号のみ表示： A-----B
結合種類も非表示： A^^^^^B

ここで，"^"は任意の結合種類を意味する．なお，実際には結合の種類も隠してしまうと，芳香環の存在が不明となる．解析者の理解容易性を考慮して，結合記号のみ表示する形式を採用した．

2.3 芳香環中原子の小文字表記

これまでのフラグメント表記では，炭素原子はすべて C で表した．しかし，例えば下記の分子 I の側鎖部から生成されるフラグメント C3-C3-N3H, C3-C3=O1 を考えてみよう．これらフラグメントの最初の炭素原子が芳香環を構成する原子であるか否か，この表記からはわからない．



I

芳香環の構成原子であることは，元素の違いにも匹敵する重要な情報であるため，環の芳香性の認識を行い，芳香環中の原子をすべて小文字で表記できるようにした．この結果，上記フラグメントは，c3-C3-N3H, c3-C3=O1 と表され，解析者が理解しやすくなった．

2.4 カルボニル基を単一の原子として出力

カルボニル基（-CO-）は分子の化学的性質に大きな影響を与える．しかも，N 原子と隣接するとアミド基（-CO-NH-），O 原子と隣接するとエステル（-CO-O-）のように，隣接原子と組み合わさって双方の原子の性格を大きく変える．

フラグメント表現を厳密に線形に限定すると，カルボニル基の枝分かれのため，前記 I の側鎖部分は c3-C3-N3H, c3-C3=O1, O1=C3-N3H のように，３種のフラグメントで表さざるを得ない．しかし，これでは元来のアミド基が直接認識できず，ルールの読者は c3-C3-N3H の N 原子がオレフィンに結合したアミンか，アミドか迷うこととなる．

そこでカルボニル基を２配位の元素 CO として扱い，上記側鎖を c3-CO2-N3H と表して，解析者の理解を容易とした．

3. 記述子の選択

ドーパミンアンタゴニスト 1,349 種を対象とし，フラグメントの最大長を８とした場合には，8041 種のフラグメントが生成される．しかし，この記述子数は，現在のカスケードモデルの実装システムで取り扱うには多過ぎ，またごく少数の化合物でのみ現れるフラグメントを多数含むため，そのすべてを利用することは現実的とは言えない．そこで，化合物中での出現頻度が 15%-85%（*edge* = 0.15 と表記する）の範囲にある 114 フラグメントのみを選択し，解析に用いてきた．

結果として，出力されるルールの条件部には芳香環，アミン構造や酸素原子など，ドーパミン活性に寄与するフラグメントが現れ，活性に寄与する構造的特徴を判断することができた．しかし，ルールを支持する化合物構造の視察を行った結果，より支持度の低いフラグメントが活性に重要な寄与をする場合があると判断できた．

他方，フラグメントの中には O1=S4-c3:c3H と S4-c3:c3H のように，一方が他方の部分構造でありながら同じ出現頻度を持ち，同一化合物群に現れることが自明な対が多数存在した．そこで，取り扱う記述子数を増加させるため，これら相関の高い冗長な記述子を削除することにより，より低頻度の記述子を解析に含めることを試みた．

3.1 名義変数間の相関係数定義

Gini は名義変数での分散定義を以下のように行っている [4]．彼は，数値変数 x_i の分散が次式で表されることを示した．

$$V_{ii} = \left(\sum_a (x_{ia} - \bar{x}_i)^2 \right) \Big/ n = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2 \quad (1)$$

ここで，V_{ii} はi番目の分散であり，x_{ia} は a 番目の事例における x_i の値を，n は事例数を示す．

ついで，(2)式の距離定義を(1)式に代入すれば，名義変数間の分散が(3)式で定義できる．これが Gini-index としてよく知られた量である．

$$x_{ia} - x_{ib} \begin{cases} = 1 & \text{if } x_{ia} \neq x_{ib} \\ = 0 & \text{if } x_{ia} = x_{ib} \end{cases}, \quad (2)$$

$$V_{ii} = \frac{1}{2n^2}\sum_a\sum_b (x_{ia} - x_{ib})^2 = \frac{1}{2}\left(1 - \sum_r p_i(r)^2\right) \quad . \textbf{(3)}$$

この定義の $(x_{ia} - x_{ib})^2$ を $(x_{ia} - x_{ib})(x_{ja} - x_{jb})$ に変更して共分散を定義しても，整合性の取れた値は得られない．そこで，値の差にベクトル表記を導入することにより，著者らは合理的な共分散の定義を導いた [5]. 値 r, s を取る変数 x_i と値 u, v を取る変数 x_j について，分散，共分散と相関係数は下の (4) - (7) 式で与えられる．

$$V_{ii} = n_{r\cdot}n_{s\cdot}/n^2 = \tfrac{1}{2}\left(1 - (n_{r\cdot}/n)^2 - (n_{s\cdot}/n)^2\right) \quad . \quad \textbf{(4)}$$

$$V_{jj} = n_{u}n_{v}/n^2 = \tfrac{1}{2}\left(1 - (n_{u}/n)^2 - (n_{v}/n)^2\right) \quad . \quad \textbf{(5)}$$

$$V_{ij} = \frac{|n_{ru}n_{sv} - n_{rv}n_{su}|}{n^2} \quad . \quad \textbf{(6)}$$

$$R_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}} \quad . \quad \textbf{(7)}$$

ここで, (6)式の分子が変数間の相関の強さに直接影響する項である．実際，完全に相関（無相関）する場合，この相関係数は1.0 (0.0) の値をとる．

3.2 相関係数による変数選択

変数選択のアルゴリズムを下に記す．

1. すべての変数対x_i, x_jについて相関係数を計算し，$R_{ij} > min\text{-}R_{ij}$.の条件を満たす時にそれらの変数対を*pairs* に追加する．
2. *pairs* をR_{ij}の降順にソートする．
3. *pairs*の先頭の変数対から x_i, x_j を取り出す．
4. もし，双方の変数が現在の変数群$\{x\}$に含まれているなら，いずれかの変数を削除する．
5. ステップ3, 4 をすべての*pair*につき繰り返す．

なお，変数を削除する場合，フラグメント名が長い方を残すことを標準としている．しかし，短いフラグメントが化学的に意味のある場合や，相関が高くともその双方が解釈に必要な場合もある．そこで，保存／削除する変数名を別途指定する機能も付加した．

図 1 は，6 種の *edge* 値に対し， $min\text{-}R_{ij}$ 値を 1.0, 0.99, 0.97, 0.95, 0.90, 0.85, 0.80, 0.75, および 0.70 と変えた場合に，選択された変数の数を対数尺度で示したものである．ここで $min\text{-}R_{ij} = 1.0$ では変数選択が全く行われず，また $min\text{-}R_{ij} = 0.99$ では完全な相関を持つ変数のみが削除されている．

この図から，*edge* 値を変更してもカーブの形は同じであること，また右端での急激な落ち込みから, およそ 20-30%の変数群が完全な相関を持っていることが分かる. さらに $min\text{-}R_{ij} = 0.90$ として変数選択を行えば，変数の数はほぼ半減する．

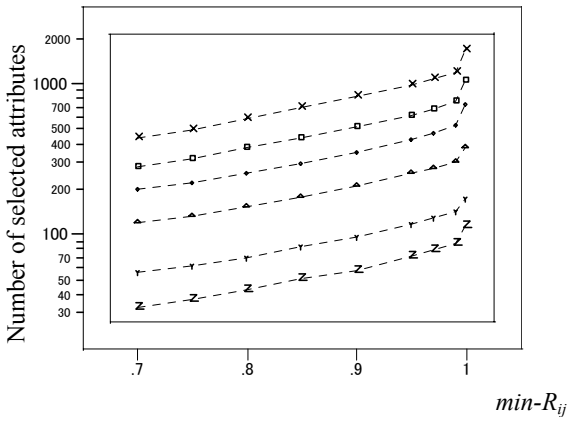


Fig. 1. Numbers of selected attributes changing $min\text{-}R_{ij}$ value.

3.3 ラティスサイズへの影響

カスケードモデルにおけるルール探索のためのラティス展開は，パラメータ*thres*で制御される. この応用ではこれまで, 100-150変数を使用し，*thres* 値は0.15-0.20に設定して, 5,000 – 30,000のラティス節点を探索してきた．

図３は探索した節点数の等高線図である．ここで, y軸は$min\text{-}R_{ij}$, x軸は (A) 変数の数(*#attributes*), (B) *edge* 値を表しており，*thres* = 0.15, 0.175, 0.20 に対する結果を示した．最も少ない節点数(*#nodes* = 3000)の等高線には矢印を付した．

(A)の等高線は, $min\text{-}R_{ij}$ が大きな値をとる領域で y 軸に平行であり，小さな値になるほど右下隅へと流れている．この結果は，相関の低い変数群を使う場合，変数数が増えてもラティスのサイズがそれほど増加しないことを示している．実際この実験では，およそ 1000 の変数から選んだ 400-500 変数を用いた計算が可能となった．

(B)では, 等高線が右上隅から左下隅への対角線に沿って描かれている．*thres*=0.15 の図中，右下から２番目の等高線近辺に位置する４つの点についての計算結果を，次の表１に示す．

Table 1. Calaculated results for 4 points near a gray contour (*thres*=0.15)

Point	*edge*	$min\text{-}R_{ij}$	*#atrs*	*#nodes*	*#detected*	*#rules*
P1	0.02	0.70	287	4992	23	6 (3)
P2	0.05	0.80	155	5983	39	8 (4)
P3	0.10	0.90	130	5223	72	9 (4)
P4	0.15	0.99	88	6265	97	14 (5)

この表から，変数の数（*#atrs*）は88 – 287と異なるが，ほぼ5,000前後の節点数（*#nodes*）が探索されていることが分かる．すなわち，相関の少ない変数群はラティスサイズを大幅に抑制することが判明した．

なお，この表には節点数に加えて，ラティス中での有意なリンク数（*#detected*）と，それを最適

化して得られたルール数（*#rules*）も示した．さらに図 3 で◇で囲んで表示した点は，識別力の高いルール群が得られたと考えられる点である．

これらの結果を総合すると，相関の高い変数を除去することにより，支持度の低い多くの変数を利用する計算が可能となり，質の高い知識発見に有効であることが判明した．

4. おわりに

本稿では，記述子表記の改良と相関係数による変数群の選択について述べた．実際の解析においては，どのようなパラメータ値で最適な結果が得られるかについての試行錯誤が必要であった．また，自動的な変数選択のみに頼ると，化学的に意味ある変数を削除することも多いため，変数選択過程を化学者の視点から制御する必要がある．

現在，ドーパミン関連活性を有する化合物群についての解析は最終段階に至りつつある．出力ルールの選択ユーティリティーの整備や，D2 受容体に対する CO 基関与の明確化等の課題への対応も含めて，解析を進めている．これまでの改良により，ルール群から活性部分構造を推定する作業は，まさにクロスワードパズルを解くような知的に興味深い作業になりつつある．発表当日には，ドーパミン関連活性に対する実際のルールや，それから推論された活性部分構造も展示する予定である．

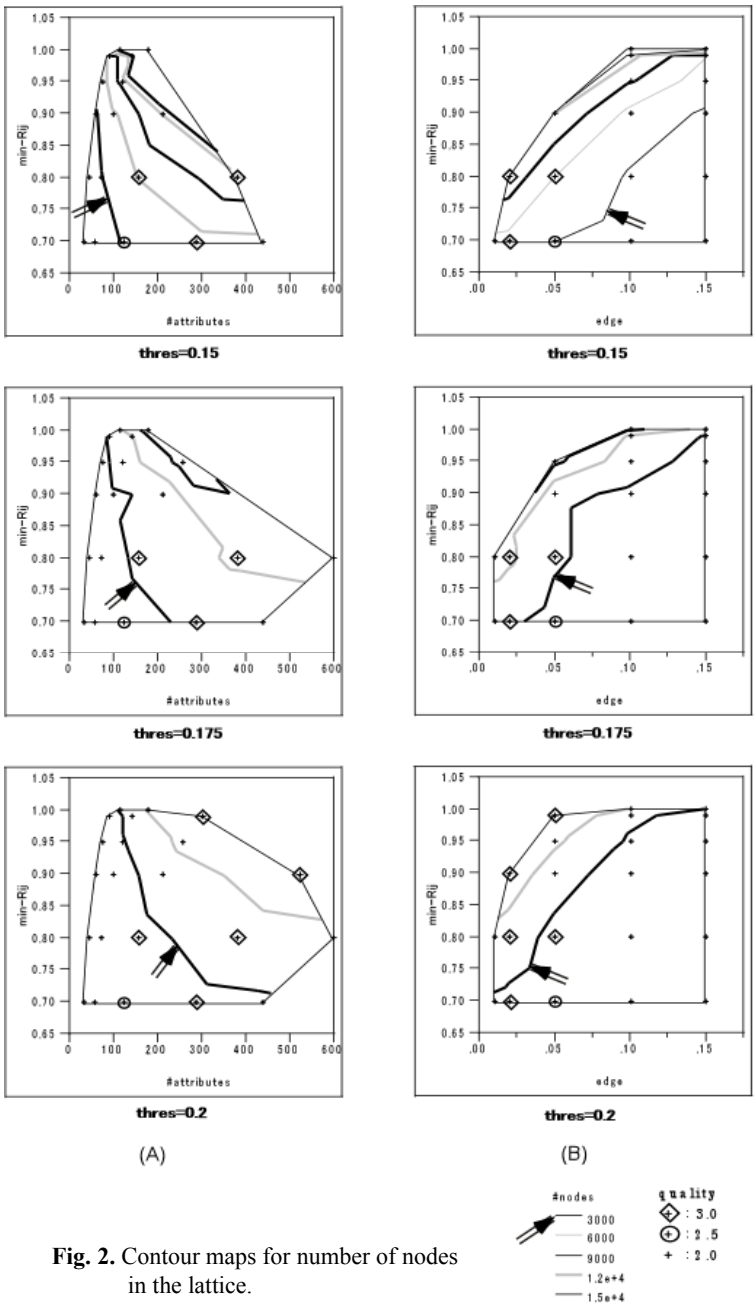


Fig. 2. Contour maps for number of nodes in the lattice.

参考文献

[1] 岡田孝、山川真透: "カスケードモデルによるリード化合物特徴のマイニング", 第 30 回構造活性相関シンポジウム, K15, pp.49-52，豊橋 (2002).

[2] 上口尚美、山川眞透、新妻弘崇、岡田孝: "ドーパミン受容体リガンドの構造的特徴について" 第26回構造活性相関シンポジウム，講演番号 K12，星薬科大学 (2003).

[3] Okada, T.: Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds. *J. Computer Aided Chemistry*, vol. 2, 790-86, 2001.

[4] Gini, C.W.: Variability and Mutability, contribution to the study of statistical distributions and relations, *Studi Economico-Giuridici della R. Universita de Cagliari* (1912). Reviewed in: Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data. *J. Amer. Stat. Assoc.* 66 (1971) 534-544.

[5] Okada, T.: A Note on Covariances for Categorical Data. In: Leung, K.S. et al (eds.) Intelligent Data Engineering and Automated Learning - IDEAL 2000. LNCS 1983, Springer-Verlag (2000) 150-157.